CS550 Project - Skin Lesion Analysis Towards Melanoma Detection

Oğuzhan Çalıkkasap
21801131
o.calikkasap@bilkent.edu.tr

Vahid Naghashi
21801136
vahid.naghashi@bilkent.edu.tr

Elmira Khajei
21801053
elmira.khajei@bilkent.edu.tr

ABSTRACT

In this report we describe our solution for 2018 ISIC challenge for lesion analysis. In this task, participants are required to recognize 7 different skin diseases from dermoscopic images. Main challenge of this dataset that it is multi-class classification, quite imbalanced and have relatively insufficient samples to train deep neural networks. We tackle this problem by applying data augmentation techniques and transfer learning. Combining with image augmentation preprocessing step, we experimented with VGG16, ResNet50 and InceptionV3 models as we fine-tuned them in order to increase class-based accuracies which is the most critical part of this challenge. We trained our networks on the original dataset besides the augmented one and consolidated the contributions of our approach.

1. INTRODUCTION

Skin cancer is a major public health problem, with over 5,000,000 newly diagnosed cases in the United States every year. Melanoma is the deadliest form of skin cancer, responsible for an overwhelming majority of skin cancer deaths. In 2015, the global incidence of melanoma was estimated to be over 350,000 cases, with almost 60,000 deaths. Although the mortality is significant, when detected early, melanoma survival exceeds 95% [1]. The goal of this project is to develop an intelligent image analysis tool to enable the automated diagnosis of melanoma from dermoscopic images.

2. DATASET

Our data was extracted from the ISIC 2018 (ISIC 2018: Skin Lesion Analysis Towards Melanoma Detection) grand challenge datasets. Training data for the last challenge consisted of 10015 RGB lesion images that have the size of 450x600 and associated file with labels for seven disease categories which are Melanoma (MEL), Melanocytic nevus (NV), Basal cell carcinoma (BCC), Actinic keratosis / Bowens disease- intraepithelial carcinoma (AKIEC), Benign keratosis- solar lentigo / seborrheic keratosis / lichen planus-like keratosis (BKN), Dermatofibroma (DF) and Vascular lesion (VASC). Additionally, to that task, we were given supplemental information about images in the dataset. This data included image identifier with associated lesion identifier and diagnosis confirm type. Images with the same lesion identifier showed the alike primary lesion on a patient, as mentioned on challenge forum. There are a couple of drawbacks of this dataset as it is highly imbalanced and contains a considerable number of noisy images. Class-based number of instances are shown in table 1. The noise can be described as the reflections of measurement units or the scope borders on the images which cause inconsistent lesion imaging. A sample of noisy image is demonstrated in figure 1, which has a measure scale on top and scope borders in corners. On the other hand, since challenge committee does not provide validation and test labels, we had to split training set into train, validation and test sets. Since classification scores are originally tested on 1157 images, we chose the same instance numbers as our validation and test set numbers.

Table1: Number of instances for each disease class

MEL	NV	BCC	AKIEC	BKN	DF	VASC
1113	6704	514	327	1099	115	142

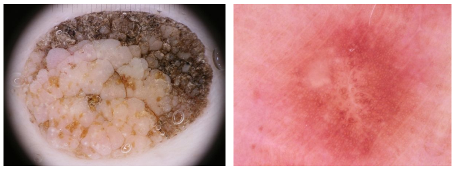

Figure 1. Noisy and normal image samples from the dataset

3. DATA PREPROCESSING

3.1 Image Augmentation

As mentioned earlier, isic archive contains a highly imbalanced dataset. Difference between two classes, NV and DF, reaches up to 67-folds, which is quite a challenging task to learn a decent classifier for each class. To prevent network from a biased classification towards a dominant class, we have performed several image augmentation techniques to have a balanced dataset as it directly affects class-based accuracy scores which is critical for this task. To this end, we have have augmented images from fewer-instanced classes with 15 different combinations of random rotation, updown flip, left-right flip, add random noise, contrast change, brightness change, sigmoid correction, and swirling operations. We specified a certain range for changing the parameters of each of these operations and have them randomly selected in this range. For example our rotation angle is chosen from a +45 -45 degrees interval and swirling strength is from 0 to 3 in magnitude. An example of augmentation operations on a single image is shown in figure 2.

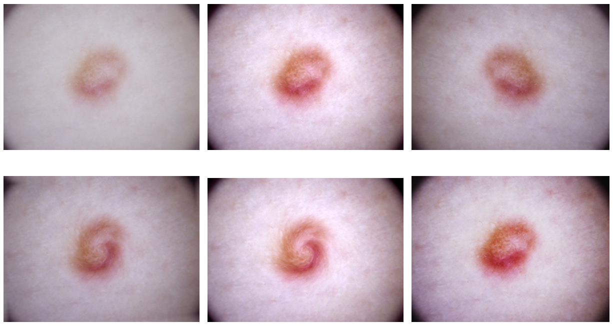

Figure 2. Brightness, contrast, sigmoid, flip and swirl augmentation

After generating enough samples of classes which have fewer instances, we finalized our augmented train dataset, merging with the original one. On the other hand, downsampled images from the dominant class NV. Finally, we ended up with an augmented dataset of 3k instances for each class, forming up a total of approximately 21k images and assigned them in different folders for the classifier input.

4. MODELS

When we train a network from scratch, we encounter some limitations such as the following. First, huge data is required for a decent training. Since a deep network has a lot of parameters, we need to have a numerous data in order to get an optimal set of parameters. Second, huge computing power is required. Even if we have enough data, training generally requires multiple iterations and requires quite a computational resource.

Fine-tuning a network is to modify the parameters of an already trained network so that it adapts to the new task at hand [2]. Knowing that the initial layers learn very general features and as we go higher up the network, the layers tend to learn patterns more specific to the task it is being trained on. With this in mind, we wanted to keep the initial layers intact and retrain the later layers for fine-tuning our task of melanoma classification. This helped us to avoid limitations such as the amount of data required for training and the time required to train the network. Hence, we benefit from not training the entire network. Additionally, the part that is being trained is not trained from scratch. Since the parameters that need to be updated is less, the amount of time needed will also be less.

4.1 Deep Residual Networks

Deep residual networks (ResNets) [3] has achieved state-of-the-art results in image classification and detection related problems. Compared with many deep networks, e.g., VGGNet, adding extra layers (beyond certain depth) results in higher training and validation errors. Therefore, it is challenging to optimize a very deep networks with many layers. ResNets architecture consists of a number of residual blocks with each block comprising of several convolution layers, batch normalization layers and ReLU layers. The residual block enables to bypass (shortcut) a few convolution layers at a time. Therefore, the ResNets architecture is capable to overcome this limitation by adding shortcut connections that are aggregated with the output of the convolution layers. In this ISIC 2018 skin lesion analysis challenge, we propose to exploit the deep ResNets for robust visual features learning and representations.

4.2 VGG16 Network

VGG16 (also called OxfordNet) is a convolutional neural network architecture named after the Visual Geometry Group from Oxford [4], who developed it.The VGG network is characterized by its simplicity, using only 3x3 convolutional layers stacked on the top of each other in increasing depth. Reducing volume size is handled by max pooling. Two fully-connected layers, each with 4,096 nodes are then followed by a softmax classifier. In this work we freezed all the layers of the network except the last four layers and trained the network in that fashion.

4.3 Inception v3 network

Inception network was once considered a state-of-the-art deep learning architecture (or model) for solving image recognition and detection problems. These networks are invented to avoid overfitting and increasing number of parameters issues. It's main idea is based on increasing the width of the network rather than its depth [5]. Here, we exploited Inception v3 in order to avoid overfitting problem and getting better accuracy. One layer of Inception network is shown in figure. 1 in which 1x1 convolutions are used to reduce dimensionality. In our project, we exploited transfer learning by using pre-trained Inception network. We kept all the layers' weights constant except that we again trained the last four layers, including two fully-connected, dropout and softmax layers of Inception model. By this way we used the merits of transfer learning and speeded up the training process.

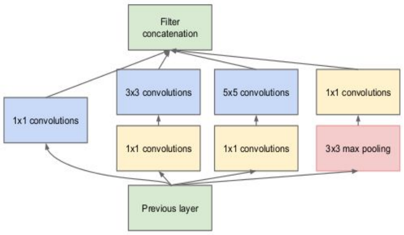

Figure 3. Single Inception layer

5. EXPERIMENTS

We have conducted our experiments on original and augmented datasets using three different models to see the effect of different network configurations as well as the dataset characteristics. Datasets like ISIC2018 are considerably challenging for classifier networks due to their imbalanced and non-standardized imaging nature, so even state-of-the-art classifiers and CNNs cannot handle this issue and they are prone to being overfitted. Cross-validation might be a good solution, but sometimes it takes a long time to train a network several times with different train data portions. To overcome this problem, we used data augmentation and also fine-tuned the networks for this specific task. We performed transfer learning in which pre-trained Imagenet weights are used for each of our models. Keeping first convolutional layer parameters fixed, we trained last four convolutional layers together with the additional fully connected layers we have added at the end of each network. These additional layers consist of 512 nodes with dropout factor of 0.5 and ReLu activation function due to our prior grid search results. These layers are added because of having our networks more specialized on our dataset by also training their weights. Stochastic batch gradient descent with learning rate of 0.001 has been used as the optimizer in all of our experiments. After several training sessions, we also decided to include a zooming augmentation and noted increase in accuracy since perspective of the images were varying in such manner as well.

Input image size for VGG16 and ResNet is 224*224*3, and 229*229*3 for Inception networks. Therefore, we re-sized the images before feeding them to our models. Then, pixel values are normalized to the range between 0 and 1. Results of different models have been compared based on accuracy, recall, F1 measure. In addition, we have calculated class-based accuracy for each model. The results corresponding to each network are illustrated in more detail in the relevant subsections.

6. RESULTS AND DISCUSSION

We have determined optimal network parameters after experiments for our task as shown in the table 2.

Table2: Network parameters

	VGG16	ResNet50	InceptionV3
Learning rate	0.001	0.01	0.001
Decay	1e-6	1e-8	1e-6
Momentum	0.9	0.9	0.9
Batch size	64	100	64

We discuss our experiment results on training and validation accuracy graphs, relying on validation results to see the real performance during learning. Then we provide class based scores improved by our work and briefly give a conclusion to our study.

6.1 Imbalanced Dataset Results

VGG16:

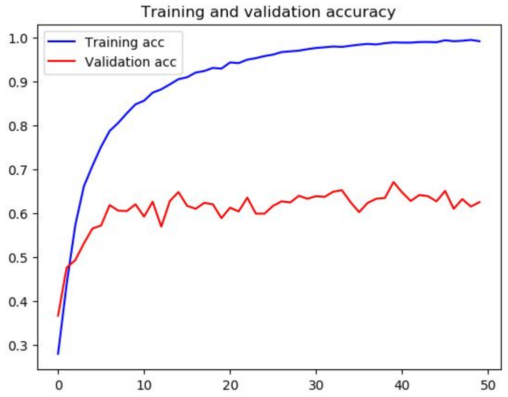

Figure 4. Imbalanced train results of VGG16

We observed an overfitting in this experiment since training accuracy reaches up to 99%, by contrast, validation accuracy stuck only at around 60%.

ResNet50:

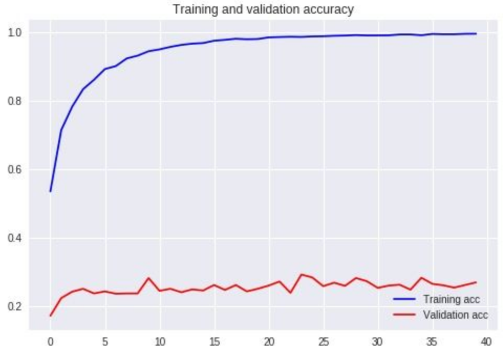

Figure 5. Imbalanced train results of ResNet50

Similar to the VGG model, we observed an overfitting in this Resnet training, with accuracy up to 99% on train set, while, validation accuracy reaches only to 30% which is even lower than VGG case.

InceptionV3:

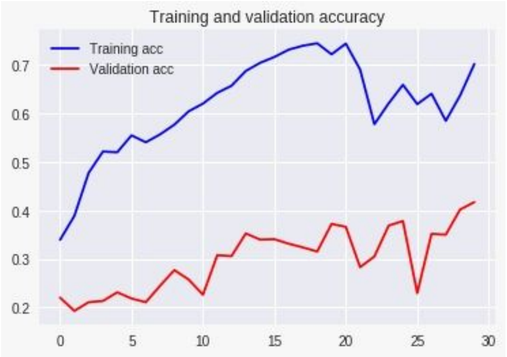

Figure 6. Imbalanced train results of Inception

Because of lack of enough train data and imbalanced class-based data distribution, Inception model doesn't give satisfactory results on both train and validation data sets. . In Inception model in spite of depth, width of the network is extended by using different convolutions, such as 1*1, 3*3 and 5*5 convolutions [6]. We trained the last layers of this network and kept the other layers freezed. The other layers has been trained using ImageNet database and we used this trained network. But, again because of size of parameters and lack of enough data, the Inception v3 model became overfitted to train set and its validation accuracy is relatively low. In ResNet50 model all layers were trained using ISIC2018 balanced train dataset and we expected a high validation accuracy. But, because of the network size and huge number of parameters, this model has also become overfitted.

6.2 Augmented Dataset Results

Compared to the accuracy graphs of imbalanced trained networks in the previous part, we observed quite satisfactory accuracy improvement in validation set along with the train set accuracy. They increase in parallel as seen in the figures and that makes more sense in terms of a reliable network, in other words not overfitted.

VGG16:

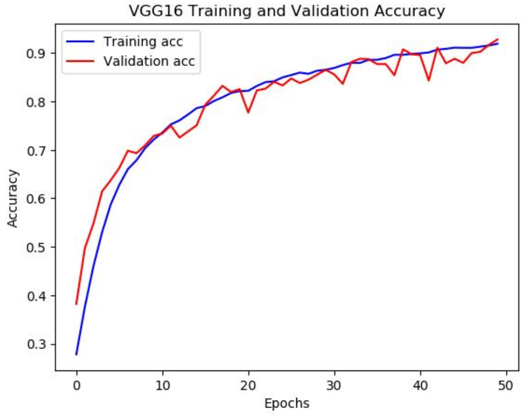

Figure 7. Augmented train results of VGG16

Resnet50:

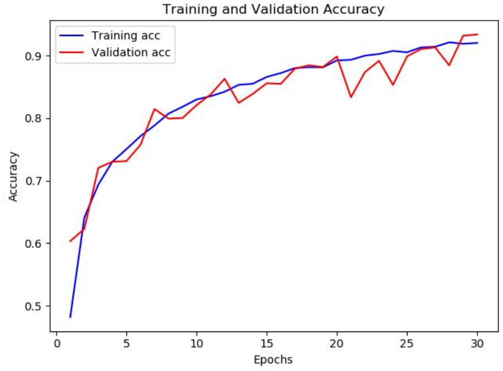

Figure 8. Augmented train results of ResNet50

We observed a promising accuracy even at the first epoch and ResNet50 reached a slightly better performance than VGG16 network. We also observed that ResNet converges faster compared to plain counter part of it.

InceptionV3:

As we expected, InceptionV3 model gave better results after data augmentation, in terms of both train and validation datasets. Balancing the number of images corresponding to each class led to better performance of this well-known network.

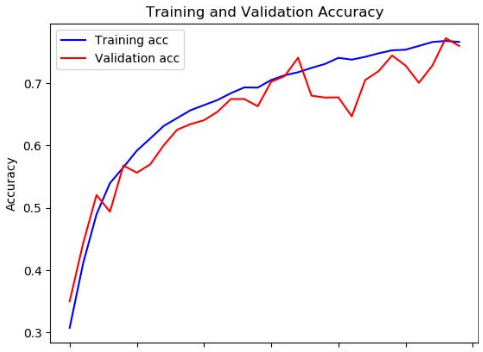

Figure 9. Augmented train results of Inception

7. COMPARISON OF CLASS-BASED PERFORMANCE

Here we compare the contribution of our work by going only over the results of VGG16 model to be more direct and brief. We evaluated our models by the precision, recall and f1-score metrics to be more precise to observe how our approach improved class-based accuracies as well as the overall accuracy. While state-of-the-art overall accuracy for this challenge is around 87%, we achieved an overall 93% accuracy, keeping a decent class-based scores as shown in figure 11. We also compared the same model with original dataset in order to see the effect of our augmentation techniques and there is a considerable improvement in class accuracies especially with fewer instances.

	precision	recall	f1-score	support
0	0.94	0.83	0.87	191
1	0.95	1.00	0.97	182
2	0.88	1.00	0.88	175
3	0.95	1.00	0.89	173
4	0.89	1.00	0.95	221
5	0.97	0.68	0.79	105
6	0.95	0.92	0.90	133
micro avg	0.92	0.92	0.90	1180
macro avg	0.93	0.91	0.89	1180
weighted avg	0.93	0.92	0.90	1180

Figure 10. Imbalanced train classification report of VGG16

	precision	recall	f1-score	support
0	0.94	0.79	0.87	191
1	0.93	0.93	0.93	182
2	0.84	0.96	0.90	175
3	0.96	0.90	0.93	173
4	0.90	0.98	0.94	221
5	0.98	0.95	0.97	105
6	0.98	1.00	0.99	133
micro avg	0.93	0.93	0.93	1180
macro avg	0.93	0.93	0.93	1180
weighted avg	0.93	0.93	0.93	1180

Figure 11. Augmented train classification report of VGG16

8. CONCLUSION

Biomedical image classification task has become one of the hot research fields recently and different challenges are open in this field. ISIC 2018 challenge is one of them and third task of this competition is to classify seven different dermoscopic images of skin diseases. For this purpose, three different and state-of-the-art deep learning models: VGG16, ResNet50 and Inception version 3, have been used. These deep CNN models have been widely used in various computer vision tasks. We exploited these models to classify the images from ISIC2018 task 3 dataset. VGG model with 16 layers and filters of size 3*3 gave the best result. As we expected this network with relatively lower number of parameters gave the best validation accuracy. This number of parameters and many convolutions embedded in this architecture helps to avoid overfitting and at the same time to learn efficiently from train data, leading to a good performance on the validation set. On the other hand, other deeper models with 50 and 42 layers had a poor performance on the test phase. This shows that complex models are sensitive to imbalanced dataset problem, while shallower networks, such as VGG16 perform better. As a result, providing enough and balanced data set is one of the critical prerequisites for gaining satisfactory results in machine learning projects.

9. REFERENCES

[1] International Skin Imaging Collaboration (ISIC) (2018). *About Melanoma*. Retrieved from https://challenge2018.isic-archive.com/

[2] Gkioxari, Georgia, et al. "R-cnns for pose estimation and action detection." arXiv preprint arXiv:1406.5212 (2014).

[3] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[4] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[5] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[6] Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.